Journal of CHEMOMETRICS

ONLINE DISCRIMINATION OF CHEMICAL SUBSTANCES USING STANDOFF LASER INDUCED FLUORESCENCE SIGNALS

Journal:	*Journal of Chemometrics*
Manuscript ID	Draft
Wiley - Manuscript type:	Special Issue - Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Kraus, Marian; DLR, TP-APW Gebert, Florian; DLR, TP-APW Pargmann, Carsten; DLR, TP-APW Walter, Arne; DLR, TP-APW Duschek, Frank; DLR, TP-APW
Keyword:	standoff identification of chemical agents, classification algorithms, laser induced fluorescence, machine learning

ONLINE DISCRIMINATION OF CHEMICAL SUBSTANCES USING STANDOFF LASER INDUCED FLUORESCENCE SIGNALS

Abstract

Chemical contamination of objects and surfaces, caused by accident or on purpose, is a common security issue. Immediate countermeasures depend on the class of risk and consequently on the characteristics of the substances. Laser based standoff detection techniques can help to provide information about the thread without direct contact of humans to the hazardous materials. This article explains data acquisition and classification procedure for laser induced fluorescence spectra of several chemical agents. The substances are excited from a distance of 3.5 m by laser pulses of two UV wavelengths (266 and 355 nm) with less than 0.1 mJ per laser pulse and a repetition rate of 100 Hz. Each pair of simultaneously emitted laser pulses is separated using an optical delay line[1]. Every measurement consists of a dataset of 100 spectra per wavelength containing the signal intensities in the spectral range from 250 nm to 680 nm, recorded by a 32 channel photo multiplying tube array. Based on this dataset three classification algorithms are trained which can distinguish the samples by their single spectra with an accuracy of over 95 %. These predictive models, generated with decision trees, support vector machines and neural networks, can identify all agents (e.g. benzaldehyde, isoproturon and piperine) within the current set of substances.

Keywords: standoff identification of chemical agents, classification algorithms, laser induced fluorescence, machine learning

Short Abstract

This article explains the data acquisition and classification procedure for laser induced fluorescence spectra of twenty chemical agents using laser pulses of two UV wavelengths. Three classification algorithms are trained which can distinguish the samples by their single spectra with an accuracy of over 95 %.

Introduction

When people or buildings are chemically contaminated, time is a valuable factor for the success of first responders. A fast determination of the hazardous substances is essential for an initialization of specific counter measures. In situ analyses save transportation time but often need access to the hotspot. Avoiding this, standoff measurements are more secure for operators but in general less sensitive concerning distinctive results. Laser induced fluorescence (LIF) provides the advantage of distant assays and might be reliable enough if two (or more) excitation wavelengths are used, gaining more information from the examined sample[2,3,4].

The following contribution describes a modular experimental setup and the discrimination of several chemicals by their LIF signals with three methods of machine learning.

Experimental

Depending on the scenario there can be many kinds of pollution and background materials which are worth being identified. Within this work the discrimination of an example subset of 20 different substances is described which represent four groups of chemicals. As shown in Table 1 the examined materials can be divided into different groups: fuels, lubricants, pesticides and solvents. Liquid samples are measured in pure condition and solids are dissolved in water or in diethyl ether, if required. All probes are filled in colloidal 3.5 ml glass cuvettes, excited 3.5 m apart and stirred during the measurements.

The laser system in Figure 1 was described in detail elsewhere[1], so only a short summary of the main components is presented here. The LIF is excited by frequency converted laser pulses of a Nd:YAG laser (Innolas Picolo Magna EVO III) with wavelengths of 266 and 355 nm, a pulse widths of 0.7 ns (FWHM) and a repetition rate of 100 pulses per second. These conditions enable pulse energies up to 60 mJ required for larger distances. The energies are adjusted between a few 10 nJ up to 200 µJ depending on the intensity of the fluorescence signal using a polarizer and a half waveplate. To gain independent excitation the 355 nm pulse is delayed by approximately 100 ns to the 266 nm one by multiple reflections between two mirrors which are slightly tilted with respect to each other. Afterwards the optical paths are united and guided onto the target. The resulting fluorescence signals are collected by an off-axis parabolic mirror with a diameter of 10.2 cm (Edmund Optics #83-957) focusing the radiation onto the input facet of an optical fiber that guides the radiation to a grating based spectrometer (Hamamatsu A10766). The radiation is diffracted onto a photomultiplier tube (PMT) array with 32 channels spanning a spectral range from 250 to 680 nm. For the electronic signals of each excitation an integration time of 50 ns is achieved by a high speed data acquisition system (Vertilon PhotoniQ).

Measurements are performed indoors and each dataset consists of 100 background corrected LIF signals of both excitation wavelengths. Every substance is measured five times which results in an overall dataset of 10000 labeled spectra with 64 features for the following data processing.

Data preprocessing

The spectroscopic signals below the excitation wavelengths are eliminated as well as the channel in the spectral region of the notch filter which blocks the second excitation wavelength in the first signal. Raman peaks may occur in adjacent regions to the excitation

wavelength. Within a wavenumber range between 3200-3600 cm^{-1} (peak positions at 292 nm and 405 nm, respectively), water causes one of the largest shifts to be expected in this set of substances (symmetric and asymmetric stretch vibrations). The concerned features are channels 3 and 4 in the 266 nm signal and channels 9 to 12 in the 355 nm signal and are eliminated. Furthermore, some spectra of low intensity show a slight elevation around 532 nm, which is caused by a remaining signal of the second harmonic output of the laser. So for both excitation wavelengths the according two channels are also set to zero. For reasons of comparability the spectra are scaled by setting the maxima to 1 and the minima to 0. The resulting dataset with the remaining 43 relevant channels is visualized for four substances in Figure 1.

Classification

Developing a classification model mainly consists of two steps called training and test which are performed with two parts of a dataset. Here, 75 % of the data are used as training set to learn how to distinguish the spectra as good as possible according to their origin. Afterwards, the remaining test set is classified to evaluate the goodness of fit of the previously learned model according to its accuracy, i.e. the proportion of correctly and totally predicted spectra.

The discrimination of the substances is investigated by making use of three different classification methods. The decision tree algorithm *C5.0*[5,6] extracts distinctive features which separate the spectra and calculates optimal benchmarks for its decisions. The so constructed borders between the intensity values of different signals are rearranged by minimizing the distances. Finally, the tree is pruned by eliminating redundant branches. Support vector machines[7] (SVM) map the data onto a hyperspace where two respective features of the signals can be separated linearly considering only adjacent points of different classes which are called support vectors. The final model is created by a simultaneous mapping of all channels, a division of all substances and a subsequent back transformation. The visualization of a small subspace separating two of the substances is shown in Figure 2. Artificial neural networks consist of multiple linear combinations of features which are linearly combined in one or more hidden layers. The coefficients (or weights) of every combination is the result of iteration and backpropagation and is explained in several articles and books[6,8,9].

For this dataset and the three mentioned methods, each model results in an overall accuracy of approximately 99 % in classifying the test set. Within a measurement containing 100 signals there are at most four single spectra which are misclassified. The confusion matrix of the *C5.0* model, averaging 25 developed trees, applied on the test set is shown in Table 2 with an accuracy of 99.2 %. The SVM model has an accuracy of 98.8 % using the radial basis function as kernel[6]. With a preceding principal component analysis[10] and one hidden layer with twelve units the neural network reaches 99.3 %.

Additional resampling methods like cross-validation and bootstrapping in the partitioning step as well as in the training are included to ensure less over-fitted and more universal valid models. All described computations are operated on a desktop PC (Intel Xeon E5-1630 v4, 3.7 GHz, 32 GB RAM) using *R* version 3.4.4[11] and mainly the packages *C50*[5], *kernlab*[7], *nnet*[8] and *caret*[12]. The latter enables resampled partitioning and the use of several classification algorithms within a common data structure.

Summary and Outlook

The results show that LIF spectra of several chemicals can be separated with different methods of machine learning techniques. All of the presented algorithms distinguish the spectra with a very good performance of approximately 99 %. For an online execution a trained model will be implemented in the setup after the data acquisition system gaining a prediction of the measured substance within a few seconds. With a maximum of four misclassifications per measurement consisting of 100 spectra, the online discrimination of chemical substances seems feasible.

Present and future experiments will be performed outdoor at distances up to 130 m on a laser test range operated by the German Aerospace Center (DLR) in Lampoldshausen, Germany. For a promising recognition of the examined materials new measurements will be investigated with lower concentrations as well as various backgrounds or mixtures of different ratios and these data have to be taken into account for the modeling process. Due to expected additional influences from atmosphere, a combination of miscellaneous algorithms or a multilevel classification might be the key to extend the limits of detection, to low error rate and to high sensitivity.

References

1. Gebert F, Kraus M, Fellner L, et al. Novel standoff detection system for the classification of chemical and biological hazardous substances combining temporal and spectral laser-induced fluorescence techniques. *The European Physical Journal Plus*. 2018;**133**(7). doi:10.1140/epjp/i2018-12147-2.

2. Duschek F, Fellner L, Gebert F, et al. Standoff Detection and Classification of Bacteria by Multi-spectral Laser-Induced Fluorescence. *Advanced Optical Technologies*. 2017;**6**(2):75-83.

3. Sivaprakasam V, Lin H-B, Huston AL, Eversole JD. Spectral characterization of biological aerosol particles using two-wavelength excited laser-induced fluorescence and elastic scattering measurements. *Optics Express*. 2011;**19**(7):6191. doi:10.1364/oe.19.006191.

4. Kraus M, Fellner L, Gebert F, Pargmann C, Walter A, Duschek F. Classification of Substances Combining Standoff Laser Induced Fluorescence and Machine Learning. *Light & Laser Current Trends*. 2018;**1**(3). http://www.heraldopenaccess.us/fulltext/Light-&-Laser-Current-Trends/Classification-of-Substances-Combining-Standoff-Laser-Induced-Fluorescence-and-Machine-Learning.php.

5. Kuhn M, Weston S, Coulter N, R. Quinlan MCC code for C5.0 by. *C50C5.0DecisionTrees Rule-Based Models*.; 2015. https://CRAN.R-project.org/package=C50.

6. Kjell Johnson MK. *AppliedPredictiveModeling*. Springer New York; 2013. https://www.ebook.de/de/product/20211095/kjell_johnson_max_kuhn_applied_predictive_modeling.html.

7. Karatzoglou A, Smola A, Hornik K, Zeileis A. kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software*. 2004;**11**(9):1-20. http://www.jstatsoft.org/v11/i09/.

8. Venables WN, Ripley BD. *ModernAppliedStatistics S*. Fourth. New York: Springer; 2002. http://www.stats.ox.ac.uk/pub/MASS4.

9. Lantz B. *Machinelearning r - second edition*. PACKT PUB; 2015. https://www.ebook.de/de/product/24829302/brett_lantz_machine_learning_with_r_second_edition.html.

10. Bishop CM. *PatternRecognition Machine Learning*. Springer-Verlag New York Inc.; 2006. https://www.ebook.de/de/product/5324937/christopher_m_bishop_pattern_recognition_and_machine_learning.html.

11. Team RDC. *RALanguage EnvironmentStatisticalComputing*. Vienna, Austria: R Foundation for Statistical Computing; 2008. http://www.R-project.org.

12. Kuhn M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software, Articles*. 2008;**28**(5):1-26. doi:10.18637/jss.v028.i05.

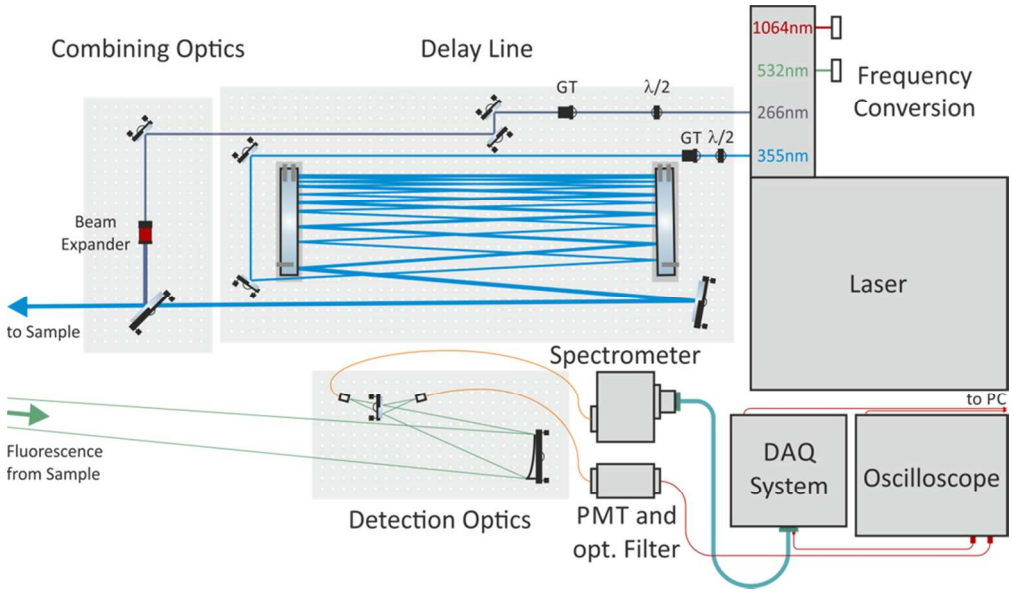
Table 1: List of substances which are measured with the current LIF setup and subsequently discriminated by their spectra. Liquids are measured pure, solids are dissolved either in water or diethyl ether depending on their solubility.

Fuel	Lubricant	Pesticide	Solvent
Diesel	Anderol555	Imidacloprid(w)	Benzaldehyde
Jet fuel	Motor oil	Isoproturon(d)	Cyclopentan
Kerosene		Malathion(w)	Ethyl acetate
Paraffin		Oxyfluorfen(d)	Isopropyl alcohol
		Permethrin(d)	Losin100
		Terbuthylazine(d)	p-Xylol
		Piperine(w)	Turpentine substitute

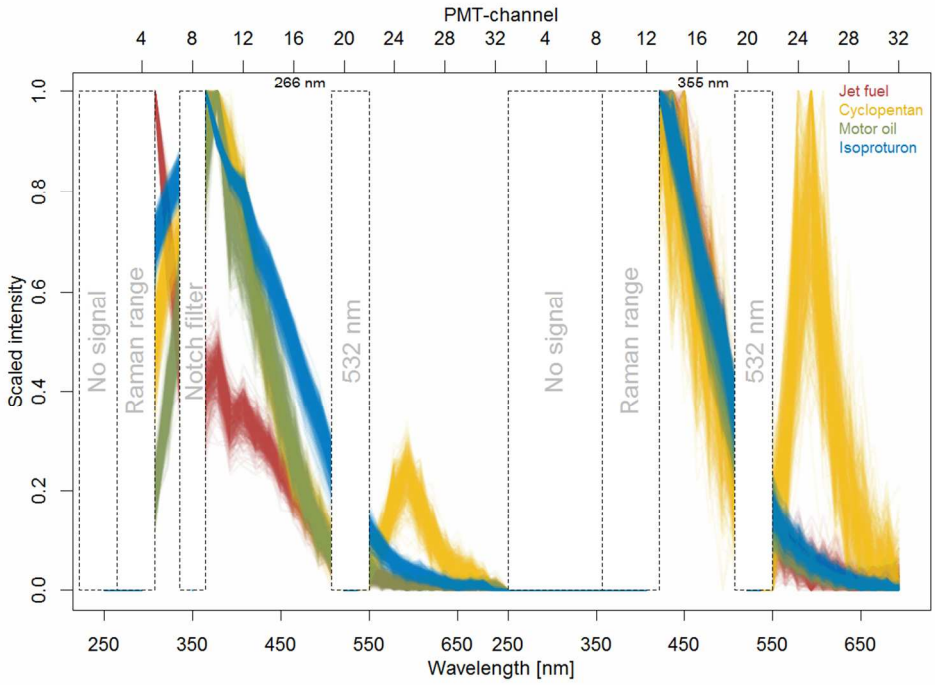
pure liquid substances; pesticides dissolved in water (w) or diethyl ether (d)

Table 2: This confusion matrix shows the predictions of *C5.0* applied on the test set with an accuracy of 99.2 %.

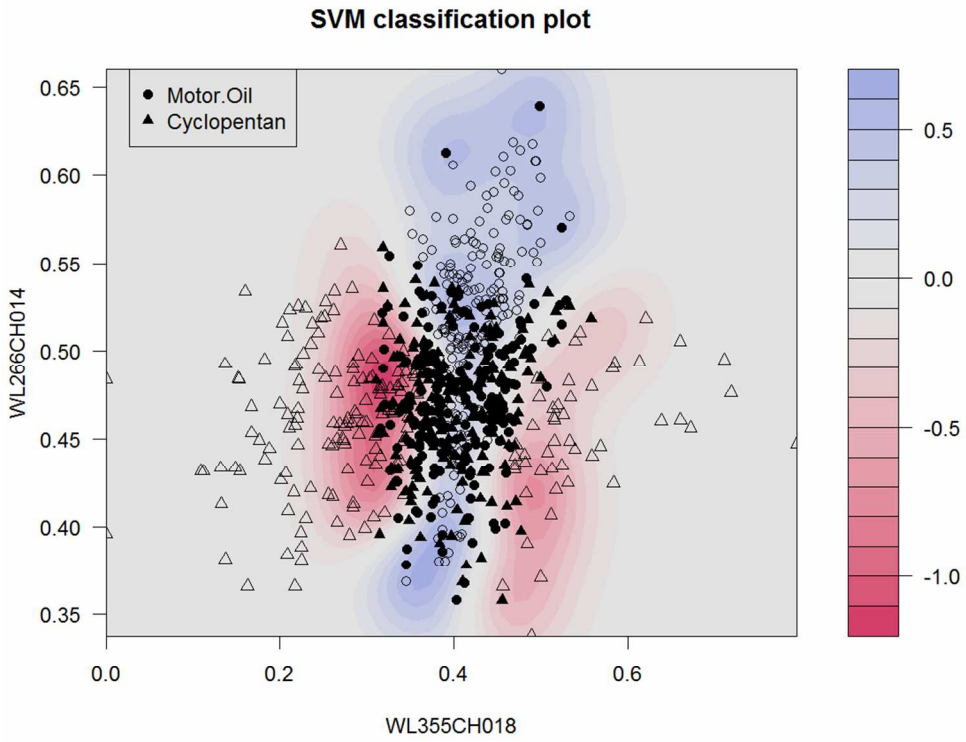
Reference / Prediction	Anderol555	Benzaldehyde	Cyclopentan	Diesel	Ethyl acetate	Imidacloprid	Isopropyl alcohol	Isoproturon	Jet fuel	Kerosene	Losin100	Malathion	Motor oil	Oxyfluorfen	p-Xylol	Paraffin	Permethrin	Piperine	Terbuthylazin	Turpentine substitute
Anderol555	125																			
Benzaldehyde		124																		
Cyclopentan			125																	
Diesel				125																
Ethyl acetate					122							2								
Imidacloprid						125														
Isopropyl alcohol							125													
Isoproturon								119									3			
Jet fuel					1				125											
Kerosene										125										
Losin100											125									
Malathion		1			2							122								
Motor oil													125							
Oxyfluorfen														123			2			
p-Xylol															124					
Paraffin															1	125				
Permethrin								6				1		2			120			
Piperine																		125		
Terbuthylazin																			125	
Turpentine substitute																				125



Schematic view of the experimental setup showing all important components, the optical paths of the excitation pulses (blue and purple) and the detectable fluorescence signal (green).



Normalized LIF spectra of four representative substances showing the regions of eliminated channels.



This SVM model separates two of the substances using only two features. In hyperspace, on which the kernel function maps to, the curvy borders are planes.